Towards a Visual Lexicon: The Creation of a Corpus Linguistic Database Using Digital Movie Data

Sebastian Sainoo-Fuller

Abstract

Corpora Linguistics have been around for a long time but have undergone a quiet revolution due to the advent of the Information Revolution. This paper looks at some issues regarding the construction and application of Corpora, and proposes the creation of a Corpus Linguistic Database using Digital Movie Data. Due to recent technological progress, searchable lexical databases can be linked to movie files with relative ease to create sophisticated concordances. By seeing lexical items in action, the learner can appreciate subtleties in nuance, body language, tone and meaning, such as for example any non-verbalized understood subtexts occurring within a dialogue. In this way, when students reproduce language many pragmalinguistic failures may be averted. The feasibility of such a project is discussed, along with the practical limitations created by proprietary law and current technology. Finally a teaching example is presented.

コーパス言語学の歴史は長く、情報革命に伴って、次第に大きな展開をとげることになった。本論文では、先ず、コーパスの構築とその適用に関する諸問題を見て、次に、デジタル映像を取り込んだコーパスの作成を提案する。最近の技術の発展により、語彙のデータベースは比較的簡単に映像情報とリンクさせることができ、高度なコンコーダンスを作成することが可能となった。これにより、学習者は、調べたい語彙に関して、その細かいニュアンスや、実際に使われている場面等を知ることができ、語用言語学的間違いを防ぐことができるようになる。さらに、本論文では、そのようなコーパスの作成の難しさを著作権等の問題から論じ、最後に、実例を紹介する。

Introduction

What is a Corpus Linguistics?

The Oxford Companion to the English Language describes a traditional corpus thus;

Usage begins 13c: from Latin corpus body. The plural is usually corpora. (1) A collection of texts, especially if complete and self-contained: the corpus of Anglo-Saxon verse. (2) Plural also corpuses. In linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database. Currently, computer corpora may store many millions of running words,

whose features can be analysed by means of tagging (the addition of identifying and classifying tags to words and other formations) and the use of concordancing programs. Corpus linguistics studies data in any such corpus... T.McA.
[from *The Oxford Companion to the English Language*, ed. McArthur & McArthur 1992, after Ball 1997]

However, as Ball (1997) suggests, in modern linguistics a corpus is taken to mean a database created from written texts which is searchable electronically, and with the results usually tabulated as a concordance. The concordance can then be reedited, or post-processed to answer specific research questions, specifically regarding, say, English usage. The most famous corpora are huge samples of spoken and written language. The well established British National Corpus (BNC) is a 100 million word collection designed to represent a wide cross-section of British English, whilst the American National Corpus (ANC) had a first installment of 10 million words scheduled for release in September 2002. The Cambridge International Corpus has been developed over the last decade and a half to help in *writing Easy* Reader teaching materials for learners of English, with the majority of data coming from magazines, newspapers, non-fiction and websites. The most famous TEFL／TESOL corpus is probably the Collin's COBUILD project, which has a 56 million word sample. Not all corpora are designed for electronic retrieval, a classic example being Project Gutenberg. Despite being the most comprehensive, accurate and most popular open source collection of digitalized paper texts and eBooks to date, the internal text of each file cannot be automatically retrieved and presented as an online concordance.

These points having been made, however, many corpora are designed for use with a rather more modest sample, very often with a single research question in mind, Ball cites Hiatt's example where a concordancing program was applied to quantify the concept of a masculine versus feminine writing style in nineteenth century American fiction. Such a corpus may be much smaller, although the raw figures themselves are somewhat hard to visualize, but Ball (1997) suggest a million words represents approximately eight medium sized books, or about five large undergraduate dissertations. So conversely 2000 words would only be about 2 pages of a dissertation or five pages of a medium sized book. Clearly such a small sample would be inadequate for most applications. A screenplay is a different matter, since it also contains stage and directions as well as raw dialogue. One example downloaded from http://www.script-o-rama.com/table.shtml is Tony Curtis's original screenplay for the 1999 movie Notting Hill. There are exactly 17,122 words, in 3672 paragraphs consuming 99377 characters including spaces. As an estimate, perhaps ten to fifteen percent of this consists of lighting and camera directions. Released in the same year, Alan Ball's screenplay of Same Mendes's Oscar winning American Beauty contains 113455 characters and spaces, broken into 2509 paragraphs, using 20,255 words. So clearly, a collection of around 70 screenplays would

make a respectable corpus of a million words. A single screenplay may well make a reasonable corpus for a specific research or educational query. McCarthy, Michael & Carter, Ronald (2001) demonstrate admirably that in a language teaching environment size need not be a significant problem.

Proposal for a Visual Corpus capable of creating a Visually Indexed Concordance

As mentioned above a text only corpus has many benefits and applications I terms of research, curriculum building as well as in a students personal development. However what is being proposed here is a radical enhancement of the concordancing facility, so that with each retrieved datum result there is also a link to the respective portion of the video file, so that the individual lexeme may be placed within a visual and aural context. The benefits of such and enhancement are manifold: at the basic level the concordancer acts as a visual search engine allowing the user to pinpoint the specific use of language from anywhere in the movie database. At a more profound level, the lexeme is placed in a social context. By seeing lexical items in action, the learner can appreciate subtleties in nuance, body language, tone and meaning, such as for example any non-verbalized understood subtexts occurring within a dialogue. In this way, when students reproduce language many so-called "cultural mistakes", or more accurately, pragmalinguistic failures, may be averted.

Under the Hood: Technical details and Feasibility analysis

Technically speaking, this is a straightforward problem. Most DVDs contain numerous files which are linked to the actual video file at runtime. These include closed caption and subtitle files, as well as other features. One possible way of creating a database would be to have searchable subtitle files, which contain only the raw dialogue spoken in the movie. It should be noted that the subtitle files are already indexed and linked to the visual files so the concordancing software need not consult external metadata to accomplish this task. However, above sample size was considered and with a typical two and a half hour long movie containing around 15,000 lexemes, a more desirable approach would be to have a database of external screenplays, as a searchable text file which could be linked to a series of video files using a MarkUp language. The concept of tagging and mark-up languages within corpora is not a new one and is discussed at length by Leech, G. (1993) in his discussion on annotated corpora and by Souter, C. (1993). However what is proposed is an inline tagging system which links directly to the chapter index on the DVD／ video file. In this way the file size can be minimalized, since only one copy of the video need be stored on the system. This is effectively a relational database which can produce a concordance form displaying a hyperlink to a specific chapter in a movie. So although a single DVD corpus can be drawn up using the subtitle files on the DVD itself, a larger corpus consisting of multiple movies can be constructed using a single relational metadatabase linked by tags to multiple video files. The other benefit of

tagging are also accessible: tags can be used to create lexemes which may ignored during a search (such as stage/camera directions or foreign words, for example). Additionally tags may be used to mark grammatical function, for instance, to differentiate between usage of the word "can" as a noun or a verb.

Given a fast web connection as well as software which can package and stream digital audiovisual content there is no technical reason why such a system could not be available on the internet. The searchable dialogue fields could be accessed through a web browser using traditional form controls, the data processing could be accomplished using a server-side script (such as MySQL or PHP), and the concordance form automatically generated using ASP technology. When the NET framework truly takes off these technologies could be seamlessly integrated even by programmer of limited experience. It should be noted that from a software and a hardware perspective, this is a entirely feasible. There is also public support: the Gutenberg Project shows that the hosting of a large corpus for public access is popularly demanded. However the Gutenberg Project would be an unsatisfactory model to take as a direct analogy due to copyright concerns.

Technically simple to actualize the corpus, there are still many limitations to overcome, many of them being proprietary. Lessig (2001, p250and p188) shows that copyright law relating to the internet and digital data storage has grown stricter in recent years. In the States it is now automatically the lifetime of the author plus seventy years (ninety five if we are talking about software produced by a company), with no effort to register demanded by statue. Lessig terms this "government-granted monopoly" over speech and he is highly critical of what he sees as the destruction of the commons and the consequences of this on the internet's ability to develop. The crux of the matter is that pretty much every modern commercial film is under one form of copyright or another, and that due to the local nature of copyright law (especially in the States) the content cannot be displayed anywhere without explicit permission. Due to the supranational nature of the Net, it remains to be seen to what degree large filmmaking corporations choose to pursue their rights through international courts. However, in a day and age when content is already readily available illicitly over the Net, it can be argued that it would be in the corporate world's best interest to preempt the development of movie corpora and use them instead as a market tool to push their products and develop their market share, especially in English-learning countries. Already many DVDs come as extended editions, containing bonus features, enhanced commentaries, photo scrapbooks and extra shots. Why not build a desktop PC search facility into every disc to market their product to language learners? When Nick Park's *Wallace and Gromit* is displaying such success as a teaching tool overseas, it may be time for other content providers to enter the market. Alternatively, access to the corpus could be restricted to subscribers or academic users. Such a model is demonstrably viable, as in the case of the Collin's COBUILD corpus. Of course the concordance need only show links to snippets of video, rather than requiring the end user

to download a whole movie, and can act as free advertising for the movies themselves. Whichever of these models one finally chooses, ultimately the corporation and the end user may benefit mutually.

There are one or two other limitations worth mentioning. Such a corpus contains only spoken language. Additionally many of the scenes we experience in the cinema are so fascinating because they are fantastic. Consider action movies: cars do not routinely explode, nor are people routinely shot. So to some extent the language is fantastic or popularistic. As a minor example which came up during the classroom practical, many popular screened weddings contain the words "I do", leading to the popular misunderstanding that this is the traditional form; whereas in legal fact the *Book of Common Prayer* establishes that the vows should read "I will" for a Church marriage to be sanctified. Saying this, the actors still speak convincing, plausible modern language even if the scenarios they act are rather fanciful. From this point of view, the dialogue is still authentic. When one considers slang, and expletives, this is an especially valid point. For a second language learner wondering when to use an expletive without committing a potentially disastrous pragmalinguistic failure, an action movie corpus would be starkly relevant. What is more the nature of the visual corpus enables the student to witness the surrounding circumstances the language is used in, so effectively enabling the students to have more understanding, choice and control of their learning environment. By showing students native speakers' behaviour, the teacher effectively creates a need, a desire in the students to emulate their language use. Such a need can only help to act as a motivating force.

Concordance in action: Example from teaching

The theme of *wedding ceremony* was taken to be the topic for several hour and half second year tertiary level classes here at the Junior College of Foreign Languages. The students were presented with the group task of writing wedding vows for two famous people. They were allowed access to the internet, the Common Book of Prayer, and were shown, *About a Boy* and *Friends* (Series 7 episode 24) *The one with Monica and Chandler's Wedding Part II.* The students were typically reticent and inhibited about expressing their own emotions so the task was deliberately structured to encourage them to use empathy skills and to project emotions onto well-known celebrities. To prevent mere passive appreciation of the audiovisual material, the students were supervised but allowed to search the DVDs using any subtitle/voiceover combination they desired to obtain useful target language. A handmade concordance for both videos was drawn up by the teacher[1]. Their final brief was that the couple were to have a non-traditional wedding, in the sense that although the format was to follow that outlined in the Book of Common Prayer, the couple wanted to write and express their own vows personally. A guide taken from www.ultimatewedding.com for native speakers hoping to design their own non-traditional vows was used as a guide. Finally the students acted out

their wedding as a drama, demonstrating their awareness of the significance of each step of the ceremony. Student feedback was positive, with many clearly enjoying the performance of the task as much as the language production aspect. The number of pragmalinguistic slips was also remarkably fewer than the text-only based approach used in the past. On a practical note, several of the students held part-time jobs in the hotel industry and felt that the topic may be of practical value to them in the future should they consider a full-time career in the hotel／ wedding industry in Japan.

Conclusion

It has been established that a Visual Corpus and online concordancer is a technically feasible project with the likelihood of a great deal of demand for native speakers and language learners alike. The legal constraints limit the degree such a concordance can be applied, but even in the case of a hand made concordance developed from a small corpus, interesting student centered tasks can be developed. A Visual Corpus can be a humanistic learning tool which encourages students to use videomatic sources in an active way. Students are able to move away from learning English for exams, to exploring real living English in an independent way. In summary, the Visual Corpus enables the student to move from a formal to a function study model through discriminate exploration of the English lexicon.

DVD resources

Notting Hill (1999) director Roger Michell, producer Duncan Kenworthy, writer Richard Curtis, Universal Studios ASIN: B00005JCA9

About a Boy (2003) director Paul Weitz, Chris Weitz. Universal Studios ASIN: B00005JL7Q

American Beauty (1999) director Sam Mendes, writer Alan Ball ASIN: B00003CWL6

The Incredible Adventures of Wallace and Gromit (2001) director Nick Parkes ASIN: B00005LC1I

フレンズ VII — セブンスシーズン DVDコレクターズセット vol.2 (2003)Warner Home Video ASIN: B000083J06

Internet Resources

Leech, G..1993 7 guidelines for writing annotated corpora : last accessed June 10th 2003 http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2maxims.htm

Ball, Catherine N. 1997. Online corpus and concordance tutorial: *last accessed June 10th 2003* http://www.georgetown.edu/faculty/ballc/corpora/

Kettemann, Bernard (1996) Concordancing in English Language Teaching: *last accessed June 10th 2003* http://www-gewi.kfunigraz.ac.at/ed/project/concord1.html

Collins **COBUILD** Project http://titania.cobuild.collins.co.uk/

CLAWS (constituent Likelihood Automatic Word-tagging System) Metadata tagging system developed by Martin Wynne

http://www.comp.lancs.ac.uk/computing/users/eiamjw/claws/claws7.html

British National Corpus http://www.natcorp.ox.ac.uk/

American National Corpus http://americannationalcorpus.org/

Oxford Text Archive (OTA) http://sable.ox.ac.uk/ota/

Linguistic Data Consortium (LDC) http://www.ldc.upenn.edu/

Project Gutenberg http://www.promo.net/pg/

Cambridge International Corpus http://uk.cambridge.org/elt/corpus/

Bibliography

Lawrence Lessig. (2001) The Future of Ideas: the fate of the Commons in a Connected World. Random House, New York.

Souter, C. (1993) *Towards a standard format for parsed corpora.* In Aarts, J., de Haan, P. and Oostdijk, N. (eds) English Language Corpora: Design, Analysis and Exploitation, 197-212. Amsterdam: Rodopi.

Stevens, Vance (1995) *Concordancing with language learners: Why? When? What?* CAELL Journal Vol 6, No. 2 pp. 2-10.

Hiatt, Mary P. (1993) Style and the 'Scribbling Women': An Empirical Analysis of Nineteenth-century American Fiction. Westport, Conn.: Greenwood Press.

Fox, Gwyneth (1998) *Using corpus data in the classroom,* In Brian Tomlinson (Ed.) Materials development in language teaching, Cambridge

Ball, Catherine N. (1993) *Did Mary Shelley write like a man? Explorations in the methodology of language and gender.* Paper presented to the Georgetown Women's Studies Research Colloquia series, December 1993.

Ball, Catherine N. (1993) *Automated text analysis: Cautionary tales.* Paper presented at the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (ACH-ALLC93), Georgetown University, June 1993.

Stevens, Vance (1991) *Classroom concordancing: Vocabulary materials derived from relevant, authentic text.* English for Specific Purposes Vol. 10, pp. 35-46.

McCarthy, Michael & Carter, Ronald (2001) *Size isn't everything: spoken English, corpus, and the classroom.* TESOL Quarterly Vol. 35, No. 2, pp. 337-340

Chomsky, N. (1964) *"Formal Discussion"*, in Bellugi, U and Brown R. (eds.) The Acquisition of Language. Monographs of the Society for Research in Child Development 29. pp 37-9

Chomsky, N. (1965) Aspects of the Theory of Syntax, Cambridge, MA: MIT Press.

Chomsky, N. (1968) Language and Mind, Harcourt Brace, New York

Leech, G. (1993) *"Corpus annotation schemes"*, Literary and Linguistic Computing 8(4): 275-81.

Leech, Geoffrey (1997) *Teaching in language corpora: a convergence,* In Gerry Knowles, Tony

Mcenery, Stephen Fligelstone, Anne Wichman, (Eds.) Teaching and language corpora. Longman pp. 1-22

Beale, A. (1987) *"Towards a distributional lexicon"*, in Garside, R., Leech G. and Sampson G. (eds) The Computational Analysis of English: A Corpus Based Approach. London: Longman.

Barlow, Michael (2002) Corpora, *concordancing, and language teaching*. Proceedings of the 2002 KAMALL International Conference. Daejon, Korea

Abercrombie, D. (1963) Studies in Phonetics and Linguistics, London: Oxford University Press.

Altenberg, Bengt & Granger, Sylviane (2001) *The grammatical and lexical patterning of make in native and non-native student writing*. Applied linguistics Vol. 22, No. 2, pp. 173-194

Aston, guy (1997) *Enriching the learning environment: corpora in* ELT, In Gerry Knowles, Tony Mcenery, Stephen Fligelstone, Anne Wichman, (Eds.) Teaching and language corpora. Longman pp. 51-66

Barlow, Michael (1992) *Using Concordance Software in Language Teaching and Research. In Shinjo,* W. et al. Proceedings of the Second International Conference on Foreign Language Education and Technology. Kasugai, Japan: LLAJ & IALL pp. 365-373

Biber, Douglas & Conrad, Susan (2001) *Corpus based research in TESOL*. TESOL Quarterly Vol. 35, No. 2, pp. 331-335

Biber, Douglas & Conrad, Susan & Reppen, Randi (1998) Corpus linguistics: investigating language structure and use. Cambridge

Conrad, Susan (2000) *Will corpus linguistics revolutionize grammar teaching in the 21st century?* TESOL Quarterly Vol. 34, pp. 548-560

Mindt, Dieter (1997) *Corpora and the teaching of English in Germany,* In Gerry Knowles, Tony Mcenery, Stephen Fligelstone, Anne Wichman, (Eds.) Teaching and language corpora. Longman pp. 40-50

Nation, I.S.P (2001) Learning vocabulary in another language. Cambridge

Schmidt, Richard (1990) *Input, interaction, attention, and awareness: the case for consciousness-raising in second language teaching*. Paper prepared for presentation at Enpuli Encontro Nacional Professores Universitarios de Lengua Inglesa, Rio de Janeiro

Sinclair, John (1998) *Corpus evidence in language description,* In Gerry Knowles, Tony Mcenery, Stephen Fligelstone, Anne Wichman, (Eds.) Teaching and language corpora. Longman pp. 27-39

Thurstun, Jennifer & Candlin, Christopher (1998) *Concordancing and the teaching of the vocabulary of academic English*. English for Specific Purposes Vol. 17, No. 3, pp. 267-280

Willis, Jane (1998) Concordances in the classroom without a computer, In Brian Tomlinson (Ed.) Materials development in language teaching, Cambridge

注

1 An automated parser or concordance was considered but due to the possible risk of copyright infringement, a handmade concordance was drawn up listing chapter indexes. Such use qualifies as fair use for academic purposes under copyright law.

E-mail : fuller@tc.nagasaki-gaigo.ac.jp